Jan Rybicki
Jagiellonian University

VISUALIZING LITERATURE: ARTISTIC STATISTICS

"It is a truth universally NOT acknowledged that" (Burrows, *Computation* 1), in a set of, say, twenty-five novels by five authors, all one needs to know to order the books by their authors (that is, if one refuses to use such childish clues as titles and names on covers) are frequencies of some thirty, fifty, hundred, or a thousand at worst, most frequent words of such a corpus.[1] The reason why this truth is universally not acknowledged is that these most frequent words rarely go beyond function words, other "non-semantic" words, or those "semantic" words, such as "man" or "time," which owe their high frequency rank to being part of frequent idioms and set phrases. While cognitive linguists might look with approval on stylometrists who base their study of literature on those "grammatical" words, the traditional literary scholar – were he or she ever persuaded to count words as part of research – would be much more interested in words that "matter": *God*, *country*, *brother*, or *love* (or various symbolic obscurations thereof). Also, while Zipf's Law tells us that thirty or fifty most frequent word types usually account for a half of a novel's (or any other text's) number of word tokens, it does nothing to explain why these very frequent words – certainly used in a less deliberate (while, perhaps, highly deterministic) way by writers – should be enough to betray those writers' authorship through style. That is, if function-word choice can be called style. But then what else should it be called, if it so well defines and/or mirrors *how* (rather than *what*) writers write.

[1] I have shamelessly stolen my opening sentence from this paraphrase of Jane Austen's most famous opening sentence from the opening sentence of the seminal monograph of stylometry, John Burrows's *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*.

Another problem here is that the series of frequencies of function words cannot be compared with the naked eye. Although it might make some marginal sense to the mainstream literary scholar that someone would wish to count the number of times Dickens uses the word *gentleman* in any of his novels (Tabata), an inventory of *the* or *said* is not only less interesting but also less feasible. And yet, for the purposes of authorial attribution, one of the main practical applications of stylometry, the former is infinitely less useful than the latter. The snag is that this requires not only arithmetic but maths and statistics, and – unless you want to employ hundreds of Dominicans like Hugh of Saint-Cher did in the 13th century – a computer that can read as well as count. And it also requires a method that has at least an empirical if not a theoretical record of making sense out of all those numbers.

One of these, and perhaps the one most commonly used by digital literary scholars, has been derived from that used in the fundamental study of *The Federalist Papers* (Mosteller and Wallace), probably the first to focus on function words, and in the above-quoted Burrows book, which used multivariate analysis of word counts. The resulting method as used here (and in other studies originating from the same Burrowsian school) is described below. It should be stated that the whole procedure is performed by a single package, "stylo" (Eder et al.), written for R, the open-source statistical programming environment.[2]

The first thing needed, of course, is a collection of texts. In the first-illustrated case in this study, this is made up of 27 highly canonical English literary texts by 11 authors. Then, all the words from all the texts are emptied into a single "bag of words," so that the number of occurrences of each word in the bag can be counted. Thus the words for the experiment are not preselected by the researcher; they are imposed on him or her by the texts themselves.

This done, the next thing is to establish a word frequency rank list for the words in the bag. For most English texts, the first five will invariably include *the*, *to*, and *and*. Then the frequency of each word from the list is counted in each individual text. This produces a list of numbers, a fragment of which might look like this (Table 1):

[2] Stylo, complete with a Manual, is freely available at https://sites.google.com/site/computationalstylistics/.

Table 1. Raw word frequencies in selected novels.

	Tenant	*Emma*	*Mansfield*	*Northanger*	*Persuasion*	*Sense*	*Professor*
the	583	536	642	678	660	567	712
to	557	540	566	479	557	569	435
and	660	504	562	492	555	482	545
of	367	442	494	503	510	494	494
i	627	329	246	274	223	277	554
a	274	322	320	328	316	289	417
in	199	225	260	270	275.	273	282
that	191	187	174	172	175	191	149
he	226	186	161	116	190	153	114
it	232	260	235	236	206	243	177
was	180	247	275	237	265	257	202
her	173	256	323	333	239	352	230
you	286	206	169	196	124	164	170

Source: own study.

This is where the real maths start. The above numbers are not really useful for any comparison, since they are raw rather than relative values. They must now be made relative to the size of each text. The easiest way to do that would be to divide each word-type frequency count by the size, in word tokens, of each text, and this was the approach applied by Burrows in his Jane Austen study; later on, however, he produced a more sophisticated formula that converted such raw word frequencies into a measure of distance (or dissimilarity) between texts. Indeed, Burrows's Delta distance became a standard in stylometry (Burrows, "Delta"). Thus, for two texts, T and T1, and for a set of n words, the distance (the degree of similarity/difference) between them is calculated as

$$\Delta(T, T_1) = \frac{1}{n}\sum_{x=1}^{n}\left|z(f_i(T) - z(f_i(T_1)\right|$$

where

$$z(f_x(T) = \frac{f_x(T) - \mu_x}{\sigma_x}$$

where, in turn,
$f_x(T)$ = raw frequency of word x in text T;
μ_x = mean frequency of word x in a collection of texts;
σ_x = standard deviation of frequency of word x.

To express this in words rather than in algebra, Delta is the mean of the absolute differences between the z-scores for a set of word-variables in a given text-group and the z-scores for the same set of word-variables in a target text. When strings of frequencies are compared between all the texts in a set, this produces another table, which contains the Delta distances between each pair of texts, or something like this (Table 2; again, only a fragment of the complete table is presented):

Table 2. Delta distances between selected novels.

	Agnes	*Pride*	*Jane*	*David*	*Mill*	*Tom*	*Clarissa*
Tenant	0.81	1.07	0.88	0.92	0.98	1.16	1.1
Emma	1.12	0.78	1.28	1.15	1.2	1.25	1.24
Sense	1.14	0.69	1.24	1.16	1.25	1.13	1.21
Professor	1.06	1.21	0.69	0.94	1	1.27	1.3
Villette	1.07	1.26	0.65	0.91	0.96	1.28	1.3
Bleak	1.09	1.18	0.92	0.55	0.87	1.21	1.17
Hard	1.16	1.25	0.96	0.65	0.91	1.26	1.25
Wuthering	1.06	1.31	0.81	0.94	1.01	1.32	1.27
Adam	1.13	1.37	0.95	0.9	0.66	1.42	1.32
Middlemarch	1.01	1.1	0.99	0.87	0.65	1.17	1.12
Joseph	1.2	1.19	1.24	1.18	1.29	0.64	1.11
Pamela	1.15	1.24	1.27	1.19	1.26	1.11	0.67
Sentimental	1.38	1.53	1.23	1.22	1.29	1.42	1.38

Source: own study.

This is somewhat more readable. It tells us, for instance, that the most-frequent-word usage in, say, *Pride and Prejudice* is closest to that in *Sense and Sensibility* (0.69) and in *Emma* (0.79); since the other Delta values for *Pride and Prejudice* are well over 1, this is usually taken as good proof – in a real attribution experiment, with one or more anonymous texts – of the authorship of the three novels by the same person. But – and this is, at long last, somewhat closer to the title of this paper – what if an attempt could be made to visualize the results in Table 2 by plotting a diagram that would show which pairs of texts come closest to each other, and then how these pairs combine into greater entities? This would be a way to classify the texts in a set according to that set's inner patterns of similarity/difference.

Now statistics can lend a helping hand. One of the ways to visualize such a system of distances is to perform what is called a Cluster Analysis: a comparison of the strings of numbers denoting the distances between the individual texts that clusters the nearest neighbours on branches of a tree diagram.

And, at least in the case of the collection in question, it does its authorship attribution quite well (Fig. 1) – and it even recognizes siblings! This result – and the 100% attributive success – has been achieved by comparing the frequencies of the 100 most frequent words, but things do not change much – in terms of that success rate – when 300 words are used (Fig. 2).[3] However, some differences can appear at higher levels of clustering: for instance, the stable Austen/Trollope branch is no longer paired with Richardson, but, instead, is joined by a larger Brontë/Dickens/Eliot branch. And while, in this particular experiment, authorship attribution has always been perfect, the more interesting side of the various graphs – the upper-level similarities between writers and groups of writers – have been much less so.

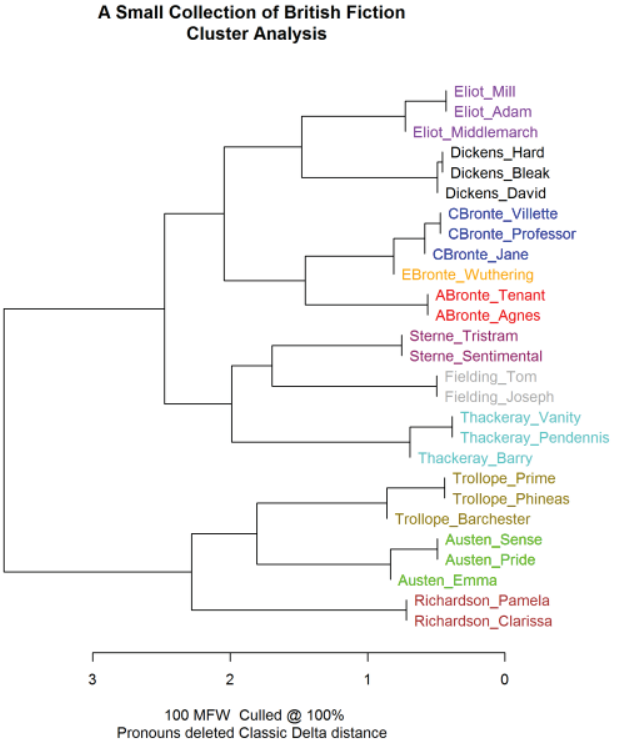

Figure 1. Cluster Analysis Tree of 27 novels based on frequencies of the 100 most frequent words.

Source: own study.

[3] The other parameters that can vary here are pronoun deletion and culling. In some (non-flexive) languages, including English, pronoun deletion improves attributive success, apparently by levelling the field between first-person and third-person narratives, or between dialogue-rich and dialogue-poor texts. Culling automatically rejects words that only appear in a certain percentage of the texts; thus, at 100% culling, only those words are used for analysis that appear in all texts in the collection at least once; at 0%, no words are rejected at all; at 50%, a word has to appear in at least half of the texts studied.

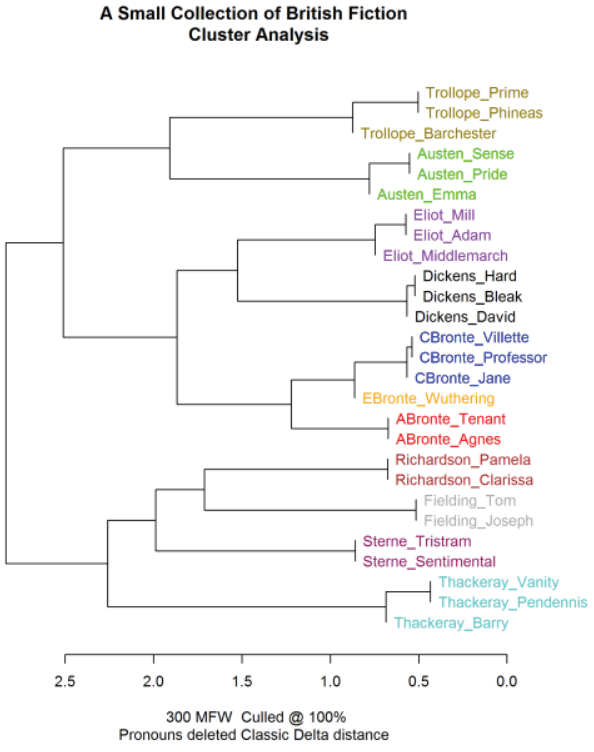

Figure 2. Cluster Analysis Tree of 27 novels based on frequencies of the 300 most frequent words. Source: own study.

There is a way around this, and a nicely democratic one at that. Since clusterings can vary slightly depending on parameters (such as reference wordlist length), why not make more runs of the Cluster Analysis and see which texts (and authors) are most frequently brought together as nearest neighbours? This may be performed with an additional procedure called Bootstrap Consensus, which does exactly that and produces another type of diagram, the Bootstrap Consensus Tree (Fig. 3), which is perhaps a more reliable product.

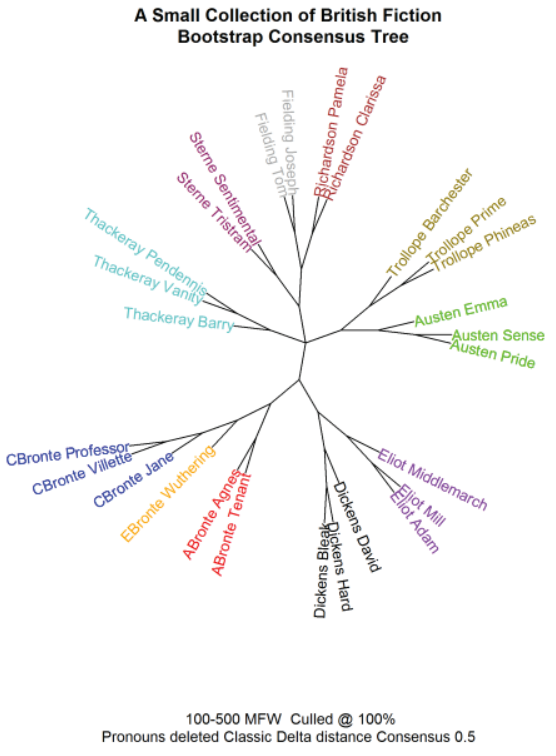

Figure 3. Bootstrap Consensus Tree of 27 novels based on frequencies of the 100–500 most frequent words.

Source: own study.

And yet… Both Cluster Analysis and Bootstrap Consensus Trees share a common drawback: they give "yea, yea; nay, nay" answers: *Mill on the Floss* is the sole nearest neighbour of *Adam Bede*, and *Middlemarch* only joins them later; Eliot's only neighbour is Dickens, etc. Such answers are rarely satisfactory in literary studies: Austen can owe as much, or almost as much, to Fielding as she does to Richardson. Thankfully, the democratic vote of the many Cluster Analyses can also be represented using network analysis, and such pieces of software as GEPHI (Bastian et al.); the result for this collection of texts is presented in Fig. 4, where the authorship attribution result is still visible; yet also visible are lesser affinities between the texts.

Other problems appear when we no longer deal with a small number of long texts (although this particular collection consists of some 6 million word-tokens). What if one would like to look for patterns of similarity that are expected by traditional histories and periodizations and classifications of literature, or perhaps for those that are not, in, say, half a thousand works? In

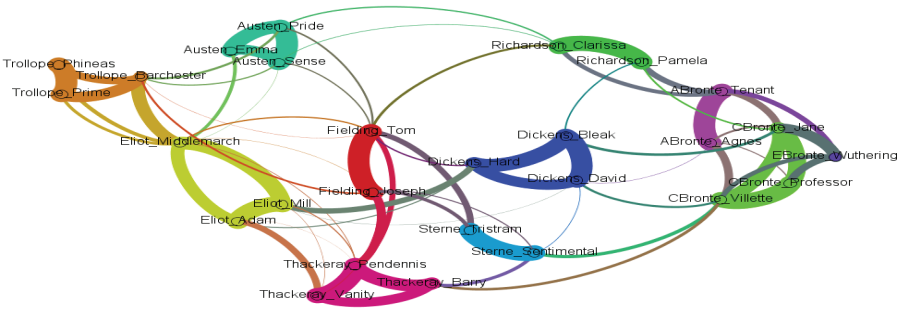

Figure 4. Network Analysis of 27 novels based on frequencies of the 100–500 most frequent words.

Source: own study.

such a case, Bootstrap Consensus Trees become unintelligible; Cluster Analysis Trees are readable but are so long, or tall, that they are only publishable on papyri. And while, for five hundred texts, a network graph would still be best viewed on a page the size of a football field, the overall effect is not unattractive (Fig. 5). The figure presents more than 500 novels in English from the beginning of the 18th century to the end of the 20th. The picture is much more complex, but certain phenomena can be observed nevertheless – above all, of course, the strength of the authorship signal. More importantly, some of the linkages make good sense in terms of standard histories of literature. The visibly separate semicircle on the North Pole of this literary globe brings together most of the 18th-century authors (both canonical and those contained in the Chawton House corpus of women writers); interestingly, these are joined, in the north-west, by some 19th-century historical romances (joined, in turn, by Tolkien's *Silmarillion*!) and, even further apart, William Morris's "prose romances." As one's eye travels south in the diagram, it encounters Peacock, Scott, and Disraeli, and then, quite chronologically, Thackeray and Dickens; mainstream Victorian women writers: the Brontës, Eliot, and Gaskell are to be found along a roughly similar latitude to the east. It should be explained that Gaskell's proximity to James is an artefact of the networking algorithm, since no direct line connects the author of *Ruth* to the author of *Roderick Hudson*, and in fact James stands mainly alone, except for one lifeline from

Defoe (*Roxana* rather than *Robinson*). The globe in the middle is a mixture of late Victorians and modernists, with a strong Joyce-Woolf connection (but she is very protean) and with a close circle, to the east, of Conrad (Polish and thus peripheral in his use of English most frequent words?) and Conrad's friends: Galsworthy and Ford are both more central, and so are the nearby Conrad/Ford collaborations. The other Slav in this English realm, Nabokov, is at least as peripheral as Conrad, but his periphery is very distant (far west) from the author of *Lord Jim*, quite in accordance with his own reaction to parallels of his and Conrad's situation: "I am too old to change Conradically" (Karlinsky 50–51). It is probably Nabokov's foreignness in his most-frequent-word usage that accounts for the fact that his nearest neighbour and his strongest connection is *Finnegans Wake*. Golding is southwest and, aptly, at *the Ends of the Earth*. Further south *sunt leones*, or bad writers: Rowling, Dan Brown and the various Ludlums and Cobens.

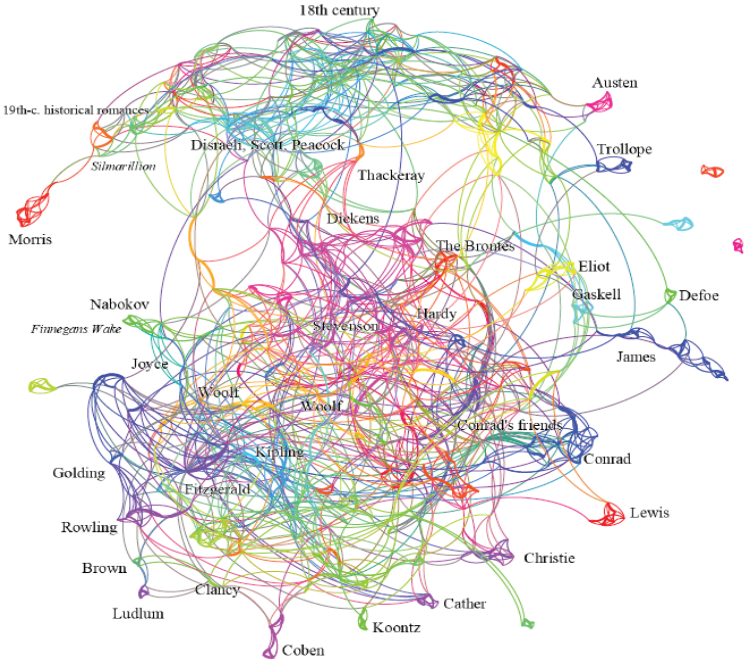

Figure 5. Network Analysis of 500 novels based on the 100–500 most frequent words.

Source: own study.

Conclusions

The reader of this text must have realized by now that she or he has been taken on an interdisciplinary rollercoaster ride, with vertiginous crests of literary fancy and abysmal troughs of down-to earth statistics. The art of literature (500 major English novels in the last example) has been reduced here to quantitative data that seemingly severed all connections between words, their combinations and their literary meaning at all levels of interpretation, thus effectively going against the usual standards of literary study; but then the data were processed by statistical methods to make some sense out of them and to return to the artistic rather than the mathematical level (diagrams of Cluster Analysis). Then the process ventured even further into the intuitive and the synthetic when the highly intricate network graph was produced for the 500 more or less representative texts of three centuries of English prose, and once again literary (and historical-literary, and intertextual) sense was imposed on the colourful image. And while this image has highly illegitimate roots (the idea of language as nothing more than "a bag of words" will be frowned upon by literary and linguistic scholars alike), the illegitimacy is somewhat mitigated by the fact that these lexical items, being very frequent and thus usually function- rather than content-words, usher in not so much single word-units of content but entire linguistic structures they represent, or at least participate in.

There are three ways in which such visualizations of literature may function. One, they are a somewhat unorthodox way of *illustrating* what we know about literature as long as they make literary sense – and very often they do, as the above examples quite clearly show. After all, if we cease to deny the powerful empirical fact that diagrams based on several dozen of very frequent words are enough to classify literary texts by their authors, it does not take much imagination to suspect that the same texts can be ordered at a higher level, i.e. beyond authorship – in a valid pattern of similarity, stylistic kinship, perhaps even intertextuality. While this phenomenon could be a result of pure luck in case of small collections of texts (such as the one networked in Fig. 4), the fact that a literarily-accepted image appears in the 500-novel diagram is a much more serious fact.

Secondly, the question appears at this point of the significance of those linkages, those lines, those clusters that make little or less sense from the traditional point of view. Should they be treated as minor errors of the clustering method, and perhaps incite the literary statistician to search for more successful algorithms? Or do they disqualify this approach to literature? Or, perhaps, do they open up new perspectives for traditional literary study? If

a particular cluster seems suspect, should we not try to re-read the texts concerned to check if, perhaps, the suspect clustering cannot be somehow explained? The medical doctor usually prefers to see lab tests of his patients before he tries to make an intuitive (and, some physicians would argue, artistic) diagnosis. Could this not serve as yet another point of view, another point of departure for quite a legitimate literary study, as legitimate as Morris Zapp's list of possible approaches to Jane Austen: "historical, biographical, rhetorical, mythical, structural, Freudian, Jungian, Marxist, existentialist, Christian, allegorical, ethical, phenomenological, archetypal, you name it" (Lodge 24). History (and literary scholarship) have since added – in earnest! – more perspectives (gender, postcolonial, queer): why not stylometric/visual?

Also, there are limits to human capacity for reading (even in the case of compulsive readers, naturally over-represented among us literary scholars). This is why digital literary historians such as Matthew Jockers call our present knowledge of periods and trends in literature "anecdotal": we have been making assumptions basing on a dozen of books each by a dozen writers, all the while ignoring thousands of other books, less canonical, less valuable, perhaps, but probably as characteristic of the same time and place (Jockers 5–10). Without ascribing to this radical view, it is nevertheless tempting to be able to produce, one day, a more robust model of, say, Victorian literature – or at least to verify the existing canonical model. And, in that case, even very traditionally-minded literary scholars might find some use for such colourful networks apart from hanging them on their (bathroom) walls.

But even if that were the only fate of this and similar images, they are, if not beautiful, then at least pretty and colourful, quite irrespectively of their classificational, interpretative, or even merely representational value. There is something exciting in the notion that all those colours and curves were generated (in an unorthodox and unusual way, it is true) solely on what those scores of writers had once written; that their individual use of that vocabulary – every little word of it – could be transmedially translated into something that can still be related to the human act of reading, understanding and interpretation – and to the pleasure of the text enhanced by the pleasure of the image.

Bibliography

Bastian, Mathieu, Sebastien Heymann, Mathieu Jacomy. "Gephi: An Open Source Software for Exploring and Manipulating Networks." *International AAAI Conference on Weblogs and Social Media*, 2009.

Burrows, John F. *Computation into Criticism: A Study of Jane Austen's Novels and an Experimenting Method*. Oxford: Clarendon Press, 1987.

——. "'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship." *Literary and Linguistic Computing* 17.3 (2002): 267–287.

Eder, Maciej, Mike Kestemont, and Jan Rybicki. "Stylometry with R: A Suite of Tools." *Digital Humanities 2013 Conference Abstracts*. Lincoln: University of Nebraska-Lincoln, 2013. 487–89.

Jockers, Matthew. *Macroanalysis. Digital Methods and Literary History*. Chicago: University of Illinois Press, 2013.

Karlinsky, Simon (ed.). *Dear Bunny, Dear Volodya: The Nabokov-Wilson Letters, 1940–1971*. Berkeley: University of California Press, 2001.

Lodge, David. *Small World. An Academic Romance*. London: Penguin, 1983.

Mosteller, Frederick. and David Wallace. *Inference and Disputed Authorship: The Federalist Papers*. Reading, MA: Addison-Wesle, 1964.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing, 2008.

Tabata, Tomoji. "More about *gentleman* in Dickens." *Digital Humanities 2009 Conference Abstracts*. College Park: University of Maryland, 2009: 270–275.